Ghoraani and Krishnan *EURASIP Journal on Advances in Signal Processing* 2012, **2012**:250
http://asp.eurasipjournals.com/content/2012/1/250

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Discriminant non-stationary signal features' clustering using hard and fuzzy cluster labeling

Behnaz Ghoraani[1,2*] and Sridhar Krishnan[2]

Abstract

Current approaches to improve the pattern recognition performance mainly focus on either extracting non-stationary and discriminant features of each class, or employing complex and nonlinear feature classifiers. However, little attention has been paid to the integration of these two approaches. Combining non-stationary feature analysis with complex feature classifiers, this article presents a novel direction to enhance the discriminatory power of pattern recognition methods. This approach, which is based on a fusion of non-stationary feature analysis with clustering techniques, proposes an algorithm to adaptively identify the feature vectors according to their importance in representing the patterns of discrimination. Non-stationary feature vectors are extracted using a non-stationary method based on time–frequency distribution and non-negative matrix factorization. The clustering algorithms including the *K*-means and self-organizing tree maps are utilized as unsupervised clustering methods followed by a supervised labeling. Two labeling methods are introduced: hard and fuzzy labeling. The article covers in detail the formulation of the proposed discriminant feature clustering method. Experiments performed with pathological speech classification, T-wave alternans evaluation from the surface electrocardiogram, audio scene analysis, and telemonitoring of Parkinson's disease problems produced desirable results. The outcome demonstrates the benefits of non-stationary feature fusion with clustering methods for complex data analysis where existing approaches do not exhibit a high performance.

Keywords: *K*-means clustering, The self-organizing tree map (SOTM), Time–frequency feature analysis, Supervised classification, Unsupervised clustering, Discriminant cluster selection

1 Introduction

The advancement in sensor technology made it possible to gather huge amounts of data, which on the one hand extends the applicability of signal analysis to a wide variety of fields, such as communications, security, biomedicine, biology, physics, finance, and geology. But on the other hand, the large data make demands for advanced and automated pattern recognition techniques to effectively process the gathered data. In pattern detection context, the general purpose of any processing technique can be described as the analysis of a given dataset to make a certain decision based on the obtained information.

In a signal classification method, a feature extraction divides a signal into short-duration segments and maps the segments into features in an appropriate multi-dimensional space. Next, a classification scheme performs the actual task of classifying the signals relying on the extracted features. In general, classification techniques can be divided into two groups: supervised learning and unsupervised learning. In a supervised learning, the classification scheme is usually based on the availability of a set of signals that have already been classified or described. Learning can also be unsupervised, in the sense that the system is not given a prior labeling of patterns. Instead, it establishes the classes based on the statistical or structural regularities of the patterns.

Supervised learning approaches are developed based on the assumption that the structures of signals from different classes are completely different. They then find a discriminating pattern among signals by dividing the feature space into non-overlapping subspaces which represent each corresponding class. Although, this approach might be satisfactory in cases the signals are separable

*Correspondence: bghoraani@ieee.org
[1]Department of Chemical and Biomedical Engineering, Rochester Institute of Technology, Rochester, NY 14623-5604, USA
[2]Department of Electrical and Computer, 1154 Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada

in the feature space, this approach seems to be too optimistic in applications where an overlap exists between different classes. This is a common issue in many real-world applications specially, in biomedical applications which the aim is to determine abnormal behaviored signals from the normal ones. In majority of cases, the discriminative structure of an abnormal signal occurs in a short duration, and as a result not the entire signal is abnormal. Hence, feature vectors that are extracted from the normal portion of an abnormal signal will overlap with the features extracted from the normal signals. In other words, natural similarities between different classes may result in some overlapping in the feature space. For example, in pathological speech recognition, while the nature of both normal and pathological signals is speech, only few high-frequency contents or transient components cause the discrimination between the two classes. Therefore, the extracted features may not necessarily represent the discriminating structures in each class, causing an overlap in the feature space. In addition, non-stationarities in the real-world signals cause some variations in the signals' properties which may result in spread and overlapping of the obtained feature vectors over the feature space.

Because of this overlapping, a supervised classifier may refuse to identify a clear discrimination among the groups and as a result may degrade the performance of the pattern recognition. Several directions have been taken in the literature to overcome the non-stationary and overlapping pattern recognition challenges as briefly mentioned in the following:

(i) Employing complex classification algorithms: complex learning methods such as artificial neural networks (ANN) [1] have been developed in order to discriminate different classes in the presence of features' overlapping.

(ii) Applying feature selection methods: there have been previous attempts to select uncorrelated feature elements that are more related to the discriminative characteristics of each class in order to improve the classification accuracy. One of these approaches is the theory of rough sets, proposed by Pawlak [2,3], is a kind of data analysis theory that introduced overlaps between classes. In this theory, a rough membership function makes it possible to distinguish similar feature elements and measures the degree of overlap between a set of experimental values representing a standard (e.g., set of values typically associated with a biomedical abnormality). This approach has been applied in feature selection and extraction to reduce a large number of features and identify the representative features [4]. It is worth mentioning that the aforementioned feature selection approaches differ from the subject of our study as the former selects the uncorrelated feature elements in a feature vector to increase the accuracy rate, while the latter keeps all the feature elements and identifies the cluster of feature vectors that are unrelated to the discrimination between classes.

(iii) Extracting the discriminant features: some attempts have been performed in the literature in order to obtain the discriminative features of the signals: local discriminant base analysis [5] and time-width versus frequency band energy mapping [6]. While these analyses are active areas of research, the optimal choice of discriminant features highly depends on the nature of the dataset and the dissimilarity measures used to distinguish between classes. Furthermore, these analyses can only be used with decomposition-based time–frequency (TF) analysis such as wavelet or matching pursuit, and are restricted to TF analysis approaches.

In an unsupervised classification method, a clustering method (e.g., Gaussian mixture model and K-means clustering) is used to obtain clusters of features for each class. This training stage is performed sequentially for each class; there is no interactions between feature vectors of different classes. In the test stage, the unknown-class data are tested with respect to the discriminant clusters of each class. The predicted class is the one associated with the clusters with the maximum probability. Unsupervised classification is a natural way to proceed towards automatic pattern recognition systems for real-world applications with overlapping features as it considers the possibility of overlapping features and clusters that share a common structure among different classes.

As our goal to enhance discriminatory powers in non-stationary feature extraction, in this study, we focus on developing a new scheme for a combined unsupervised and supervised classification approach. This framework, which we call *'discriminant cluster selection'*, aims to improve the classification accuracy in decision-making systems by providing an alternative solution to the feature overlapping problem mentioned above. In this study, we also demonstrate the fusion of non-stationary feature analysis with the proposed unsupervised classification methods to cluster the non-overlapping feature vectors as the discriminative pattern.

In this study, we employ and refine the existing clustering approaches to develop a classification technique that improves the classification accuracy rate. We adopt the notion of unsupervised clustering; however, unlike commonly used unsupervised clustering methods, we propose to perform the clustering stage on all the training feature vectors obtained from the different classes

and train one set of clusters for the entire training features. Next, we use the distribution of feature vectors in these clusters and their class label to compute the presence of the discriminative pattern in each class. Two types of clusters are identified: *discriminant clusters* which mainly consist of feature vectors from one specific class, and *common clusters* which are a mixture of features from different classes. We propose that discriminant clusters identify the representative structures in each class, and common clusters represent the similarities between classes. The proposed scheme is different from feature selection techniques which attempt to select the optimal feature elements in a feature vector to improve the classification performance. Our proposed work feeds all the elements of the feature vectors to the clustering stage, and then decides which feature cluster represents the discriminative structure between the classes. Both feature selection techniques and the proposed method can simultaneously be applied to increase the classification accuracy. In a future study, a combination of these two methods can be investigated to further improve the accuracy rate in a classification application. Our proposed framework is predicted to significantly improve the classification accuracy rate of signals. It will also improves our insight about the discrimination pattern in each class which may be reconstructed or located using the feature vectors in the discriminant clusters.

The structure of this article is as follows: Section 2 explains the discriminant feature clustering methodology. Section 3 explains K-means clustering and the self-organizing tree map (SOTM) as two unsupervised clustering techniques employed in this study. Two supervised cluster labeling techniques (hard and fuzzy labeling) are explained in Section 4. Section 5 explains the non-stationary signal features. In Section 6, the application of the developed technique is presented for three synthetic examples. In addition, the application of the proposed strategy is investigated for speech pathological detection, sudden cardiac death-risk stratification, audio scene classification, and telemonitoring of Parkinson's disease (PD), and the results are given in Section 6. Conclusion is provided in Section 7.

2 Methodology

Our goal is to identify the non-stationarity feature clusters that represent discriminative characteristics of each group. In order to proceed towards such a feature clustering approach, there is a need for a non-stationary feature extraction and clustering technique that detects the discriminant features. Figure 1 demonstrates an example to explain the proposed methodology and it's advantageous to determine such key clusters.

In this example, one normal and one abnormal signals are generated using the following equation:

$$x(t) = \sum_{i=1}^{7} x_i(t) = \sum_{i=1}^{7} \alpha_i g(\sigma_i, \mu_i) \sin(a_i t), \quad (1)$$

where $g(\sigma, \mu)$ is a Gaussian with mean μ and variance of σ^2. Mean of this Gaussian function locates a component in time, and the variance specifies the duration of each component. The sine function localizes the component in frequency domain. The normal signal is constructed to consist of seven frequency-modulated components.

To construct an abnormal synthetic signal, three of the components are transformed into transients. In many real-world applications such as biomedicine, transients are known to be the discriminative structures of abnormal signals, and are used in this example as one of the abnormality descriptors. Figure 1a–d displays the generated normal and abnormal signals in time and TF domains. In this example, spectrogram with FFT size of 1024 points and Kaiser window with parameter of 5, length of 256 samples and 220 samples overlap, was used to construct the TF of each signal. The TF domain provides TF distribution (TFD), which is a three-dimensional TF representation with two dimensions representing the time and frequency domains, respectively. The third dimension (i.e., the intensity of the distribution) indicates the energy distribution of the signal at the corresponding time and frequency. While the time representation does not provide much information about the difference between these two synthetic signals, the TFD provides a better visual display of the discriminant structure as indicated by the dashed circles. If the right quantification and classification algorithms are used, the TF representation may successfully be employed for automatic pattern recognition applications.

Six joint TF feature vectors [7] are extracted from each signal while each vector consists of three features: S_h, i.e., sparsity of the signal in time domain; S_w, i.e., sparsity of the signal in frequency domain; and D_w, i.e., abrupt changes in frequency domain. The applied TF feature extraction method is fully explained in Section 5. The extracted TF feature vectors are shown in Figure 1e. As can be seen in the feature space of Figure 1e, considering the relative location of the features in this feature space, two types of clusters can be detected: an overlapped cluster containing the frequency-modulated components which are common between two signals, and an abnormality cluster which consists of features corresponding to the transients in the abnormal signal. Our proposed feature classification method is successful if it can separate the abnormal cluster from the normal one (i.e., in this example, the transient and normal feature groups,

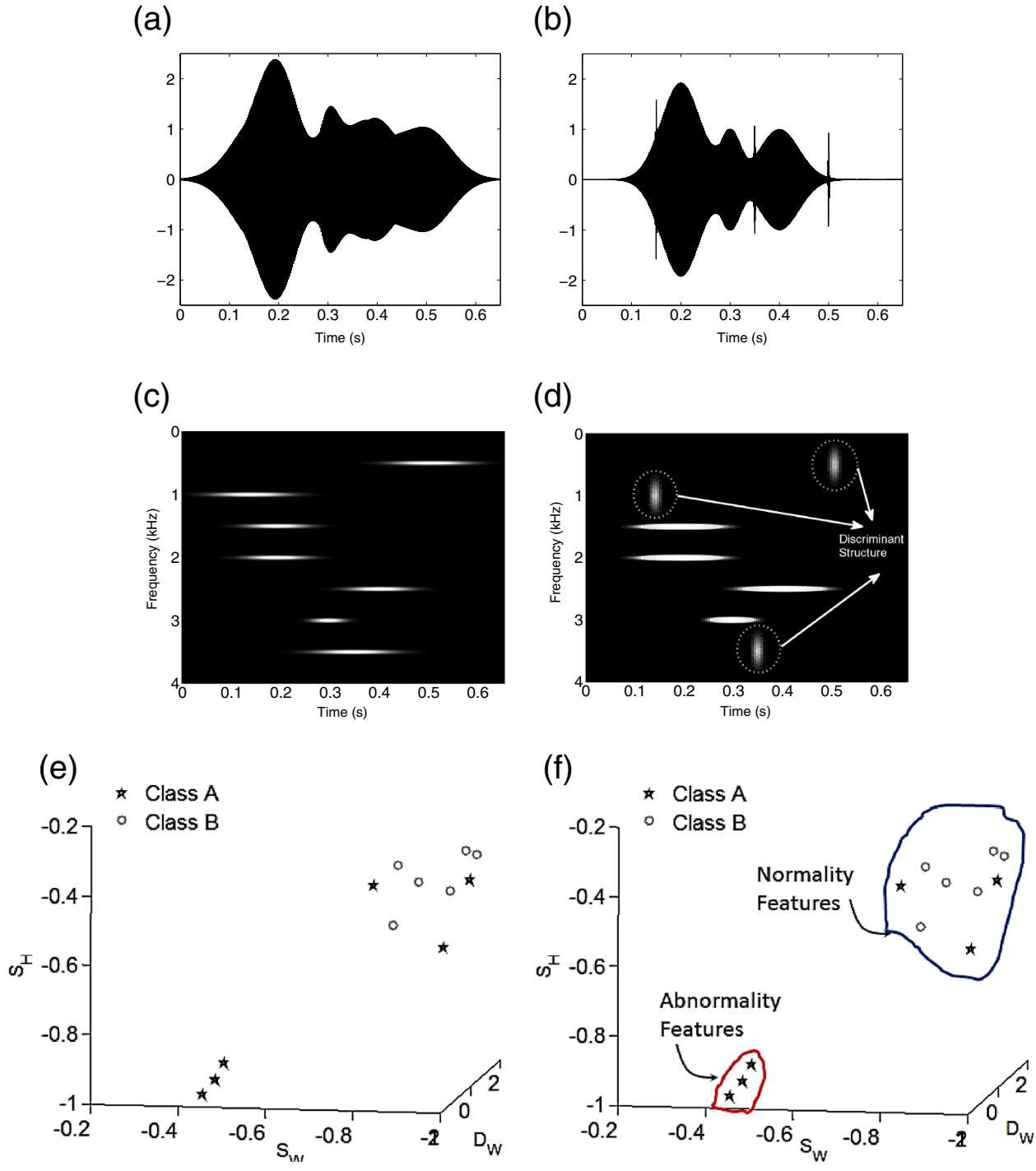

Figure 1 Discriminant feature clustering for a synthetic example. (a) Normal signal with a sampling frequency of 8 kHz is generated using Equation 1 where (α, σ, μ, a) for each component from 1 to 7 is as following: $(1, 0.005, 0.15, 2\pi 1000)$, $(1, 0.04, 0.20, 2\pi 1500)$, $(1, 0.04, 0.20, 2\pi 2000)$, $(1, 0.02, 0.30, 2\pi 3000)$, $(1, 0.05, 0.35, 2\pi 3500)$, $(1, 0.04, 0.40, 2\pi 2500)$, and $(1, 0.05, 0.50, 2\pi 500)$. **(b)** Abnormal signal is created using the following parameters for each component: $(1, 0.001, 0.15, 2\pi 1000)$, $(1, 0.04, 0.20, 2\pi 1500)$, $(1, 0.04, 0.20, 2\pi 2000)$, $(1, 0.02, 0.30, 2\pi 3000)$, $(1, 0.001, 0.35, 2\pi 3500)$, $(1, 0.04, 0.40, 2\pi 2500)$, $(1, 0.001, 0.50, 2\pi 500)$. **(c)** TF representation of the normal signal. **(d)** TF distribution of the abnormal signal. **(e)** Feature space. **(f)** Clusters representing abnormality and normality features.

respectively), and use the abnormality cluster for detection of any abnormality behavior in a test signal. The overlapped clusters do not play any role in any discrimination between the two classes. Therefore, once any feature vector is assigned to an overlapped cluster, it will be excluded from the classification of its corresponding signal, and will have no effect in labeling the signal as abnormal.

Figure 2 displays the schematic of our proposed discriminant feature clustering method. As can be seen in the block diagram of Figure 2, once TF features are extracted, a discriminant feature clustering system is introduced in order to discriminate the abnormality clusters in the feature space. This system consists of two stages: unsupervised clustering and supervised cluster labeling. In the first Stage, an unsupervised learning is performed on the entire features (i.e., both normal and abnormal) to detect all the possible feature clusters ($\{\vec{C}\}$) in the feature space. Employing this stage on the synthetic example of Figure 1 should result in two types of clusters as indicated in Figure 1f.

In the second stage, each cluster is labeled ($\{\alpha\}$) based on the feature arrangements in the feature domain determining whether the cluster consists of discriminant features or common features. The clusters which consist of the majority of abnormality signals are labeled as the discriminant structure corresponding to the abnormality pattern. The outcome of this stage in Figure 1 indicates the left cluster in Figure 1f as the abnormality cluster since all the containing features belong to the abnormal signal. Similarly, the right-hand cluster is labeled as the common cluster because the cluster consists of fairly equal number of normal and abnormal signals.

Once the abnormal and normal clusters are labeled, the trained clusters along with their labels ($\{\alpha\}$) are passed to the classification stage. In test stage, each of the test feature vectors are assigned to one of the cluster centers based on the minimum Euclidean distance (ED) measure. Next, feature vectors which belong to the overlapped clusters will be excluded, and finally, based on the membership of the test feature vectors, the class label of the corresponding signal is determined. Two methods are proposed to define the class label of each signal: *hard labeling* which is based on majority vote, and *fuzzy labeling* which is based on majority vote weighted by the membership distribution of each cluster. The above stages are described in the following Sections.

3 Clustering methods

One of the most popular clustering algorithms is K-means clustering algorithm. The other popular clustering algorithm is SOTM that does not require any information about the number of clusters in the feature domain. This Section explains the unsupervised clustering method, and the supervised cluster labeling is explained in the next Section.

3.1 *K*-means clustering

The K-means clustering is one the simplest and the most popular unsupervised clustering algorithms. The algorithm is computationally efficient and is advantageous on a dataset that consists of compact and well-separated clusters [8]. Given a set of feature vectors, $\{\vec{f}_z\}_{z=1,\ldots,Z}$,

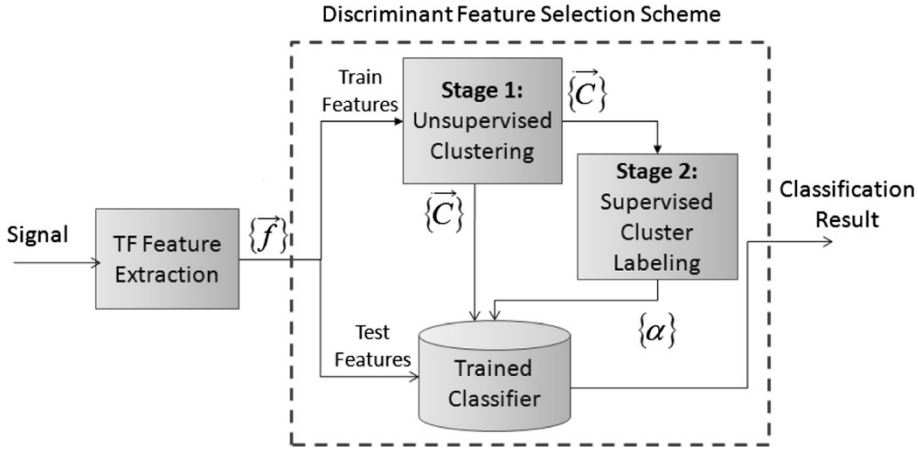

Figure 2 The block diagram of our proposed method for discriminant feature clustering. In Stage 1, a set of clusters is identified using an unsupervised clustering method which is simultaneously applied to the training feature vectors extracted from the different classes. In Stage 2, based on the distribution of the train signals in each cluster, it will be decided whether a cluster is normal or abnormal. In the test stage, the class label of each signal is assigned based on the membership of the relying feature vectors to one of these clusters.

the following phases are performed in the algorithm to identify K feature clusters [9]:

1. The method starts with K initial random centroids, $\left\{\vec{C}_u\right\}_{u=1,\ldots,K}$.
2. It classifies the feature samples into the nearest centroid according to the squared ED. To do so, it first calculates the squared ED of any given sample to all the centroids as given in the following equation:

$$\{e_z^2\} = \sum_{u=1}^{K} \left\| \vec{f}_z - \vec{C}_u \right\|^2 \tag{2}$$

Then, the algorithm assigns the sample to the centroid with minimum ED.
3. The mean of the points in each cluster is computed as the new cluster centroids:

$$\vec{C}_u = \frac{1}{Z_u} \sum_{u=1}^{Z_u} \vec{f}_z^u \tag{3}$$

where Z_u is the number of feature samples assigned to cluster u, and $\left\{\vec{f}_z^u\right\}_{z=1,\ldots,Z_u}$ are the assigned samples to cluster u.
4. The algorithm iteratively repeats Steps 2 and 3 unless the new cluster centers are the same as or close enough to the centroids of the previous Stage.

3.2 SOTM

SOTM is a type of ANN which was first introduced in [10]. The algorithm maps the data from a high dimensional Euclidean feature space onto a finite set of prototypes. Like most of the clustering algorithms, it attempts to organize unlabeled feature vectors into the clusters in a way that all the samples within a cluster are more similar to each other than those of other clusters. Each cluster is then represented using one or more prototype. Unlike classic clustering methods (like K-means) where the number of clusters should be known beforehand, in SOTM the number of clusters is determined by the algorithm based on parameters, which define the desired resolution of the clustering. The steps involved in the SOTM algorithm are briefly explained below:

1. The weight vectors are initialized randomly $\left\{\vec{C}_u(t)\right\}_{u=1,\ldots,K}$, where K is the number of clusters. The random value is usually a vector from the training set.
2. For a new input vector, the distance from the input vector and all of the existing nodes, d_u, is calculated as

$$d_u(\vec{f}, \vec{C}_u(t)) = \left\{ \sum_{z=1}^{Z} \left[\vec{f}_z - \vec{C}_u(t) \right]^2 \right\}^{1/2} \quad u = 1, \ldots, K \tag{4}$$

where $\vec{C}_u(t)$ is the node of the cluster u at time t.
3. Select the node with the minimum distance d_u as the winning node, u^*

$$d_{u*}(\vec{f}, \vec{C}_u(t)) = \min d_u(\vec{f}, \vec{C}_u(t)) \tag{5}$$

4. The minimum distance, $d_{u*}(\vec{f}, \vec{C}_u(t))$, is then compared with $H(t)$, the hierarchical control function, which decreases over time. If the input vector is within the threshold $H(t)$ of the winning node, the weight vector is updated based on the following update rule:

$$\vec{C}_u(t+1) = \vec{C}_u(t) + \lambda(t)[\vec{f} - \vec{C}_u(t)] \tag{6}$$

Where $\lambda(t)$ is the learning rate, which decreases with time. When the input vector is farther from the winning node than the threshold, a new subnode is generated from the winning node at $\vec{f}$.
5. Checking the terminating conditions; The algorithm will stop if any of the following conditions are fulfilled
   - Maximum number of iterations is reached.
   - Maximum number of clusters is reached.
   - No significant change occurs in the structure of the tree.
6. Otherwise the algorithm is repeated from Step 2.

The hierarchical control function acts as an ellipsoid of significant similarity. $H(t)$ can be assumed as a global vigilance threshold that is used for measuring the proximity of a new input sample to the nearest existing node in the network. Samples that fall outside the scope of the nearest existing node result in generation of a new node as child of the winning node. By initializing $H(t)$ to start from a large value, the clusters discovered at the early stages of the clustering will be far from each other. Decay of $H(t)$ over time results in partitioning the feature space in low resolution at the early stages of the clustering, while favoring partitioning at higher resolutions later. There are two standard hierarchical control functions proposed for the original SOTM algorithm: linear and exponential decays.

$$\begin{aligned} H(t) &= H(0) - \left[\left(1 - e^{-\zeta/\tau H} H(0)\right)/\zeta\right] t, \\ H(t) &= H(0) e^{-t/\tau H}, \end{aligned} \tag{7}$$

where τH is a time constant, which is bound to the projected size of the input feature F, $H(0)$ is the initial value, t is the number of iterations (or sample presentation), and ζ is the number of iterations over which the linear version of $H(t)$ would decay to the same level as the exponential

version. One benefit of initializing $H(t)$ to a large value, possibly larger than the maximum variation within the data, is that all levels of resolution across the data can be explored.

The learning rate in Equation (6), $\lambda(t)$, is an important factor in organizing the network. $\lambda(t)$ can operate in number of different global or local modes. In global modes, a single learning rate is applied to all nodes, whereas in local modes an individual rate operates for each node a set of nodes. There are a few modalities proposed for the operation of the learning rate and the details are discussed in [11,12].

4 Cluster labeling

Assignment of the right label to each cluster is one of the critical concerns in our proposed discriminant cluster selection system. We propose two methods to label the obtained clusters and obtain the class label of the signals as explained in the following subsections.

4.1 Method 1: hard labeling

In an E-class classification problem, this method decides whether each cluster represents classes $1, 2, \ldots$, or E.

- First, the clusters are identified, say K clusters $\left\{\vec{C}_1, \vec{C}_2, \ldots, \vec{C}_K\right\}$. $K \geq E$ is the number of clusters and is not necessarily equal to the number of classes (i.e., E). The number depends on the application and the employed clustering method.
- Next, we calculate the feature vectors of each class based on their assignment to a cluster and denote this number as $\mathrm{NUM}_1(u)$, $\mathrm{NUM}_2(u)$,..., $\mathrm{NUM}_E(u)$ representing the number of class 1 to E feature vectors in the uth cluster, respectively.
- Then, clusters with a fairly equal mix of feature vectors from different classes are identified as overlapped clusters and labeled as common clusters (i.e., K_c clusters). The remaining clusters (i.e., K_d clusters) are discriminant clusters and are labeled based on the membership distribution of their feature vectors. The class with majority membership defines the label of each discriminant cluster. In order to quantify the significance of the overlap between different classes, the clusters with more that 30% of overlap are assigned to the common clusters, and the remaining clusters are identified as the discriminant clusters. The calculation proceeds as shown in the following equation:

$$\begin{aligned} &\alpha_u = 0, \\ &\text{For } u \in \{K_c\} \\ &\alpha_u = arg\ \mathrm{Max}\left\{\mathrm{NUM_e}(u)\right\}, \\ &\text{For } u \in \{K_d\} \text{ and } e = 1, \ldots, E \end{aligned} \tag{8}$$

  where α_u is the label defined for the uth cluster, and $\alpha_u = 0$ represents a common cluster.
- Once the training stage is completed, the estimated clusters and the calculated labels denoted with $\{\alpha_1, \alpha_2, \ldots, \alpha_K\}$ are passed to the test stage.

In the testing stage, first a signal is decomposed to r feature vectors. Next, each feature vector is classified based on which cluster it belongs to. Finally, based on the label of the r feature vectors, we decide on the class label of the signal. To perform this calculation, for any new feature vector $\vec{f}_{\mathrm{test}}$, the following procedure is performed:

- $\mathrm{Cluster}_{\vec{f}_{\mathrm{test}}}$, the cluster, which each test feature belongs to, is found as the nearest cluster based on ED criterion:

$$\begin{aligned} \mathrm{Cluster}_{\vec{f}_{\mathrm{test}}} &= arg \underset{u=1,\ldots,K}{\mathrm{Min}_u} \left\{\left|\vec{f}_{\mathrm{test}} - \vec{C}_u\right|\right\}, \\ &= u_f, \end{aligned} \tag{9}$$

  where $\vec{C}_u$ is the center of each cluster constructed in the training stage.
- The label of the above cluster is assigned to each test feature, and is used to determine the class $\vec{f}_{test}$ belongs to:

$$\vec{f}_{\mathrm{test}} \in \mathrm{Class\ e} \quad \text{if} \quad \alpha_{u_f} = e, \tag{10}$$

- Once all the feature vectors in a test signal are labeled, the feature vectors that are assigned to common clusters are excluded and the labeling of the remaining feature vectors are used to classify the signal. A test signal is classified as a class e signal, if the majority of its test feature vectors (i.e., excluding the feature vectors assigned to common clusters) belong to class e.

We call this procedure '*hard labeling*' as each cluster is distinguished with one label.

4.2 Method 2: fuzzy labeling

After all the feature vectors are clustered, clusters with large overlapping (i.e., containing more than 30% overlapping feature vectors) are associated as common clusters (i.e., K_c), and the remaining clusters (i.e., K_c) which are the discriminant clusters are used in the training stage as follows:

The proposed fuzzy cluster labeling calculates the label of each feature as a membership matrix $\mathbf{M}_{K_d \times E}$, where each entry in this membership matrix, m_{ue} (which is

called a membership coefficient) indicates the probability of a vector in the cluster u belongs to class e.

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1E} \\ m_{21} & m_{22} & \cdots & m_{2E} \\ \vdots & \vdots & \vdots & \vdots \\ m_{K_d1} & m_{K_d2} & \cdots & m_{K_dE} \end{bmatrix} \tag{11}$$

where E is the number of classes and K_d is the number of discriminant clusters.

The membership coefficients are calculated based on the distribution of each class in different clusters as shown in the following equation:

$$\begin{aligned} m_{ue} &= p(\theta_{\text{Class}_e}|\text{Cluster}_u) \\ &= \frac{\text{NUM}_e(u)}{m_u} \end{aligned} \tag{12}$$

where $\text{NUM}_e(u)$ is the number of features belong to class e that exist in cluster u, and m_u is the total of features in the uth cluster. These coefficients will be used in calculation of the membership degree for each of the test vectors.

In the test stage, first a signal is decomposed to r feature vectors. Each of the feature vectors representing is assigned to one cluster centers found in the previous stage based on the minimum ED criterion (Equation 9). The feature vectors located in a common cluster is excluded. Next, we simply count the number of feature vectors that are located in each discriminant cluster and recorded the numbers as a scatter vector S. The scatter vector is defined for the remaining feature vectors as follows:

$$S = \left[s_1, s_2, \ldots, s_{K_d}\right] \tag{13}$$

where s_u is the number of the representing vectors for a test signal that fall within the uth cluster and K_d is the number of discriminant clusters.

Finally, the probability of a signal belonging to a class is calculated according to the distribution of its representing feature vectors in different clusters and can be described as

$$\Phi(e) = S \times M(:, e) \tag{14}$$

where $M(:, e)$ is the eth vector of the membership matrix, M, and the signal is labeled to belong to the class associated with the maximum value of $\Phi(e)$.

Although the advantage of the hard and fuzzy labeling is the identification of the representative clusters for each class and discriminating them from the common clusters, the method requires that each class contributes with the same number of feature vectors. Since the identification of representative clusters is based on comparing the membership of each class in the clusters, the number of normal and abnormal feature vectors should be the same in order to perform a fair comparison. The proposed solution in such scenarios is to reduce the sample size of all the classes to the sample number of the smallest classes.

5 Non-stationary signal feature extraction

Figure 3 depicts the schematic of the feature extraction technique along with the proposed clustering method.

This approach captures the TF features by applying the non-negative matrix factorization (NMF) [13] to the TFD of each signal. Spectrogram can be used as a simple TF representation. Seven features are extracted from the decomposed vectors including: $\text{MO}_{\vec{h}}^{(1)}$, $\text{MO}_{\vec{w}}^{(1)}$, $\text{S}_{\vec{h}}$, $\text{S}_{\vec{w}}$, $\text{D}_{\vec{h}}$, $\text{D}_{\vec{w}}$, and $\text{SH}_{\vec{w}}$.

5.1 NMF

NMF was performed in the middle of the 1990s under the name positive matrix factorization (PMF) [14]. In 1999, Lee and Seung [13] introduced some simple algorithms for the factorization, and demonstrated the success of the technique on some classification applications. NMF decomposes a non-negative matrix ($\mathbf{V}_{M\times N}$) as follows:

$$\begin{aligned} \mathbf{V}_{M\times N} &= \mathbf{W}_{M\times r}\mathbf{H}_{r\times N} \\ &= \sum_{i=1}^{r} \vec{w}_i \vec{h}_i^{\mathbf{T}} \end{aligned} \tag{15}$$

where r is the order of decomposition, and $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices, which are called base and coefficient matrices, respectively. NMF algorithm starts with an initial estimate for $\mathbf{W}$ and $\mathbf{H}$, and performs an iterative optimization to minimize a given cost function. Lee and Seung [15] introduce two updating algorithms using the least squares error and the Kullback–Leibler (KL) divergence as the cost functions:

$$\text{Least squares error: } \mathbf{W} \leftarrow \mathbf{W}.\frac{\mathbf{V}\mathbf{H}^T}{\mathbf{W}\mathbf{H}\mathbf{H}^T}, \quad \mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T\mathbf{W}\mathbf{H}}, \tag{16}$$

$$\text{KL divergence: } \mathbf{W} \leftarrow \mathbf{W}.\frac{\frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}\mathbf{H}^T}{\mathbf{1}.\mathbf{H}}, \quad \mathbf{H} \leftarrow \mathbf{H}.\frac{\mathbf{W}^T\frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}}{\mathbf{W}.\mathbf{1}},$$

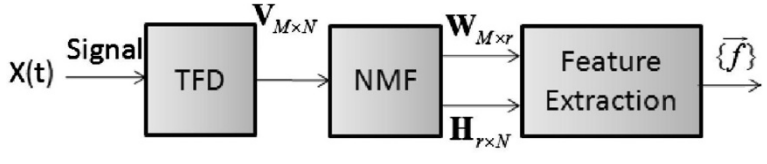

Figure 3 The block diagram of the proposed non-stationary feature extraction and discriminant classification methodology.

In these equations, $\langle A.B \rangle$ and $\frac{\langle A \rangle}{\langle B \rangle}$ are term-by-term multiplication and division of two matrices, and **1** is a matrix of ones. KL divergence formula is not a bound-constrained problem, which requires the objective function to be well defined at any point of the bounded region [16]. The log function in KL divergence formula is not well defined if any elements in matrix **V** or **WH** is zero. Hence, we do not consider KL divergence formulation in this study. The least squares error approach is a standard bound-constrained optimization problem. Various minimization strategies have been proposed for the least squares error strategy. In this study, we use a projected gradient bound-constrained optimization method which is proposed by Lin [16].

5.2 Features

As shown in Figure 3, features are extracted from each decomposed **W** and **H** matrices. The obtained features are explained as follows:

5.2.1 Joint TF moments

Moments of base and coefficient vectors (i.e., **W** and **H**, respectively) carry an important information of the TF characteristics of a signal and could be used for classification of time-varying signals [17] and feature identification [18]. We denote the ith temporal and spectral moments with $\mathrm{MO}_{\vec{h}}^{(i)}$ and $\mathrm{MO}_{\vec{w}}^{(i)}$, respectively, and compute them using the following equations:

$$\mathrm{MO}_{\vec{h}_j}^{(i)} = \mathrm{Log}_{10} \sum_{n=0}^{N} \left(n - \mu_{\vec{h}_j}\right)^{(i)} \vec{h}_j^{\mathbf{T}}(n), \tag{17}$$

$$\mathrm{MO}_{\vec{w}_j}^{(i)} = \mathrm{Log}_{10} \sum_{m=0}^{M} \left(m - \mu_{\vec{w}_j}\right)^{(i)} \vec{w}_j(m),$$

where $\mu_{\vec{h}_j}$ and $\mu_{\vec{w}_j}$ are the first moment of the jth coefficient and base vectors and are computed as follows: $\mu_{\vec{h}_j} = \sum_{n=0}^{N} n\vec{h}_j^{\mathbf{T}}(n)$ and $\mu_{\vec{w}_j} = \sum_{m=0}^{M} m\vec{w}_j(m)$.

5.2.2 Sparsity

$\mathrm{S}_{\vec{h}_j}$ and $\mathrm{S}_{\vec{w}_j}$ are the sparsity of coefficient and base vectors, respectively. These features help to distinguish between transient and continuous components. Several sparseness measures have been proposed and used in literature. We use a sparsity function as follows:

$$\mathrm{S}_{\vec{h}_j} = \mathrm{Log}_{10} \frac{\sqrt{N} - \left(\sum_{n=0}^{N} \vec{h}_j^{\mathbf{T}}(n)\right) / \sqrt{\sum_{n=0}^{N} \vec{h}_j^{\mathbf{T}}(n)^2}}{\sqrt{N} - 1}, \tag{18}$$

$$\mathrm{S}_{\vec{w}_j} = \mathrm{Log}_{10} \frac{\sqrt{M} - \left(\sum_{m=0}^{M} \vec{w}_j(m)\right) / \sqrt{\sum_{m=0}^{M} \vec{w}_j(m)^2}}{\sqrt{M} - 1},$$

The sparsity feature is zero if and only if a vector contains a single non-zero component (i.e., maximum sparsity), and is negative infinity if and only if all the components are equal (i.e., minimum sparsity).

5.2.3 Discontinuity

$\mathrm{D}_{\vec{h}}$ and $\mathrm{D}_{\vec{w}}$ represent the discontinuities and abrupt changes in each vector, respectively. These features are calculated as follows:

$$\mathrm{D}_{\vec{h}_j} = \mathrm{Log}_{10} \sum_{n=0}^{N-1} \vec{h}'^{\mathbf{T}}_j(n)^2, \tag{19}$$

$$\mathrm{D}_{\vec{w}_j} = \mathrm{Log}_{10} \sum_{m=0}^{M-1} \vec{w}'_j(m)^2, \tag{20}$$

where $\vec{h}'_j$ and $\vec{w}'_j$ are the derivatives of coefficient and base vectors as defined in the following equations:

$$\vec{h}'^{\mathbf{T}}_j(n) = \vec{h}_j^{\mathbf{T}}(n+1) - \vec{h}_j^{\mathbf{T}}(n), \quad n = 0, \ldots, N-1 \tag{21}$$

and

$$\vec{w}'_j(m) = \vec{w}_j(m+1) - \vec{w}_j(m), \quad m = 0, \ldots, M-1 \tag{22}$$

$\mathrm{D}_{\vec{h}}$ and $\mathrm{D}_{\vec{w}}$ capture the discontinuities and abrupt changes in coefficient and base vectors, respectively. A vector with a smaller value of discontinuity feature is smoother compared to a vector with a larger discontinuity feature.

5.2.4 Sharpness

$\mathrm{SH}_{\vec{w}}$ measures the spread of the components in low frequencies. In addition, we need another feature to provide an information on the energy distribution in frequency. For each base vector, first we calculate the Fourier transform as given below:

$$\vec{W}_i(\nu) = \left| \sum_{f=1}^{M} \mathrm{e}^{-j\frac{2\pi m\nu}{M}} \vec{w}_i(m) \right| \tag{23}$$

where M is length of the base vector, and $\vec{W}_i(\nu)$ is the Fourier transform of the base vector $\vec{w}_i$. Next, we perform a second Fourier transform on the base vector, and obtain $\vec{W}_i(\kappa)$ as the following:

$$\vec{W}_i(\kappa) = \left| \sum_{\nu=1}^{M/2} \mathrm{e}^{-j\frac{2\pi\nu\kappa}{M/2}} \vec{W}_i(\nu) \right| \tag{24}$$

Finally, we sum up all the values of $\left|\vec{W}(\kappa)\right|$ for κ more than m_0, where m_0 is a small number:

$$\mathrm{SH}_{\vec{w}_i} = \sum_{\kappa=m_0}^{M/4} \left|\vec{W}_i(\kappa)\right| \tag{25}$$

In order to demonstrate the behavior of feature $\mathrm{SH}_{\vec{w}}$, we assume that the base vector, $\vec{w}_i$, has two components at frequencies samples m_1 and m_2 with energies of α and β respectively:

$$\vec{w}_i(m) = \alpha\delta(m - m_1) + \beta\delta(m - m_2), \tag{26}$$

$|W(\nu)|$ (Equation (24)) is calculated as below:

$$|W(\nu)| = \sqrt{\alpha^2 + \beta^2 + 2\alpha\beta\cos\left(2\pi\ (m_1 - m_2)\ \nu\right)} \tag{27}$$

6 Results

6.1 Synthetic dataset

6.1.1 Example 1

This example is designed to demonstrate the application of the proposed discriminant cluster selection method for signal classification. We present a synthetic example of a two-class problem to demonstrate the identification process of signal classification using TF feature extraction and the proposed cluster selection method. In this experiment, we apply the TF features to a classification problem as introduced in [19,20]. Test signals are defined as the sum of two linear chirps as defined below:

$$\begin{aligned} s(t) &= \sin\left[2 * \pi\ (a_0 + a_1 t)\right] \\ &\quad + \sin\left[2 * \pi\ \left(b_0 + b_1 t + b_2 t^2\right)\right], \\ t &= 0, \ldots, N-1 \end{aligned} \tag{28}$$

where a_0, b_0 belong to a uniform distribution $U(0,1)$, $a_1 = 0.25, b_1 = 0.40$, and $N = 1024$ is the signal length. Two classes are generated by selecting b_2 from one of the following uniform distributions:

$$\text{Class1: } U\left(\frac{-0.30}{2(N-1)}, \frac{-0.20}{2(N-1)}\right) \tag{29}$$

$$\text{Class2: } U\left(\frac{-0.15}{2(N-1)}, \frac{-0.05}{2(N-1)}\right) \tag{30}$$

The TF representation for signals in each class is plotted in Figure 4.

A total number of 1,100 signals are generated in each class, and TF feature extraction and classification is performed as follows:

(i) TF representation (i.e., TF matrix) of each signal is constructed.

(ii) NMF matrix decomposition method is applied to the TF matrix, and 10 base and coefficient components (i.e., $\mathbf{W}$ and $\mathbf{H}$, and $r = 10$) are computed for each signal.

(iii) A feature vector is extracted from each component pair as explained in Section 5; i.e., there are ten feature vectors for each signal and each feature vector contains the following feature values: $\left\{\mathrm{MO}_{\vec{h}}^{(1)}, \mathrm{MO}_{\vec{w}}^{(1)}, \mathrm{MO}_{\vec{h}}^{(2)}, \mathrm{MO}_{\vec{w}}^{(2)}, \mathrm{S}_{\vec{h}}, \mathrm{S}_{\vec{w}}, \mathrm{D}_{\vec{h}}, \mathrm{D}_{\vec{w}}, \mathrm{SH}_{\vec{w}}\right\}$.

(iv) SOTM clustering is used to train and then classify the signals in each class. The classifier is trained using 90% samples and classified over all the signals. SOTM is simultaneously applied to Classes 1 and 2 feature vectors and computes 25 clusters in the feature space. The number of feature vectors associated to Class 1 or Class 2 are counted in each cluster and the distribution of feature vectors in these 25 clusters is computed and displayed in Figure 5. In both hard and fuzzy labeling, clusters with more than 30% overlap (i.e., clusters 1, 12, 13, 18, 20, 23, 24, and 25) are assigned to common clusters, and the remaining clusters are identified as discriminant clusters and are labeled depending on the labeling method proposed in Section 4 (Figure 6). In hard labeling, clusters with more than 30% Class 1 feature vectors are labeled as Class 1 (i.e., clusters 3, 6, 9, 10, 11, 15, 16, and 17) and the ones with more than 30% Class 2 feature vectors are labeled as Class 2 (i.e., clusters 2, 4, 5, 7, 8, 14, 19, 21, and 22). However, in

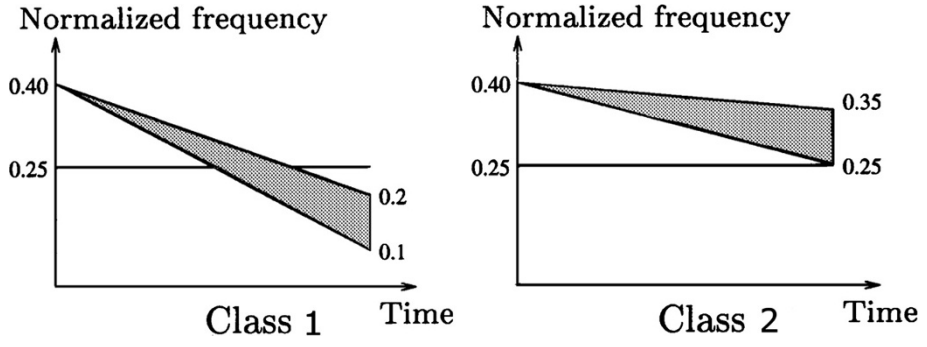

Figure 4 The gray areas represent the possible instantaneous frequencies for classes 1 and 2. This figure is courtesy of [19].

fuzzy labeling, a membership ratio is assigned to each cluster as follows:

$$\mathbf{M} = \begin{bmatrix} C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 & C_9 & C_{10} & C_{11} & C_{14} & C_{15} & C_{16} & C_{17} & C_{19} & C_{21} & C_{22} \\ 0.28 & 0.99 & 0.29 & 0.33 & 0.88 & 0.26 & 0.32 & 0.96 & 0.83 & 0.99 & 0.33 & 0.74 & 1.0 & 1.0 & 0.25 & 0.29 & 0.3193 \\ 0.72 & 0.01 & 0.71 & 0.67 & 0.12 & 0.74 & 0.68 & 0.04 & 0.1 & .011 & 0.67 & 0.26 & 0.0 & 0.0 & 0.75 & 0.71 & 0.6807 \end{bmatrix}^{\mathrm{T}} \quad (31)$$

(v) All the signals are tested and labeled. Figure 6 displays the receiver operating curve (ROC) of the final classification.

6.1.2 Example 2

The purpose of this example is to evaluate the application of the proposed method to identify an unknown discrimination pattern between two signals. Two synthetic signals (y_1 and y_2) were generated using Equation 1. Panels A and B in Figure 7 show the two synthetic signals in time and TF domains, respectively. The signals were constructed in a way that all the components, except two of them, were similar. As can be seen from the TFD plots in Figure 7, the dissimilarity components were created by transforming two of the frequency modulated components (in the right panel signal: y_1) to the linearly modulated components (in the left panel signal: y_2).

TFD in panels C and D is constructed using spectrogram method, FFT size of 1,024 points and Kaiser window with parameter of 5, length of 256 samples and 220 samples overlap. Features were extracted as explained in Section 5: NMF with a decomposition order of 10 was applied to the spectrograms of y_1 and y_2. The decomposed vectors were: $\left[\vec{w}_1(i)\ \vec{h}_1^{\mathrm{T}}(i)\right]_{i=1:10}$ and $\left[\vec{w}_2(i)\ \vec{h}_2^{\mathrm{T}}(i)\right]_{i=1:10}$, respectively. Seven TF features were extracted from each decomposed vector. Three of these features are shown in panel C of Figure 7 where the asterisk and circle correspond to y_1 and y_2 signals, respectively. K-means clustering with three clusters was applied to all the features. Each cluster with the majority membership of a signal was marked as the corresponding signal's discriminant pattern.

As can be seen in this feature plane, there was a group of features which were clustered in the middle. Using the discriminant feature selection method, this cluster was selected as the discriminant pattern in signal y_2: D_{y2}. The same method identified the discriminant pattern in y_1 signal: D_{y1}. The remaining features belonged to the commonalities between these two signals.

Panel D in Figure 7 displays the discriminant structures in y_1 and y_2 signals. These TF structures were built using the decomposed vectors corresponding to the D_{y1} and D_{y2} feature points: $\sum_{i=D_{y1}} \vec{w}_1(i)\vec{h}_1^{\mathrm{T}}(i)$ and $\sum_{i=D_{y2}} \vec{w}_2(i)\vec{h}_2^{\mathrm{T}}(i)$. As demonstrated in this example, the proposed method was able to successfully identify the discriminant structures in each signal. Once the discriminant clusters are selected, these clusters along with the proposed cluster labeling methods can be used to classify a new signal. The above example used only one signal from each class in arriving at the differences between TF structures. In practice, we have to use more number of signals in both classes before arriving at a robust discriminant pattern.

6.1.3 Example 3

This experiment introduces more challenges to the identification of the discriminant structures between two classes. In this example, the discriminant structure overlaps with the common structure; i.e., the abnormal components are mixed with the normal structure. As is demonstrated in this example, the proposed discriminant

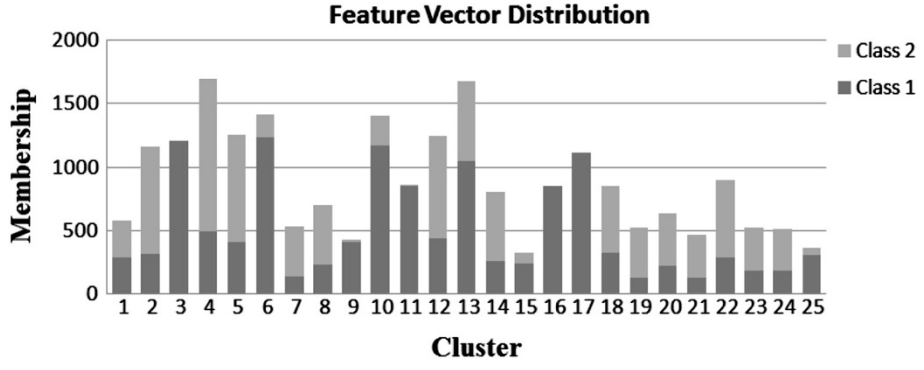

Figure 5 Distribution of Classes 1 and 2 feature vectors in 25 clusters.

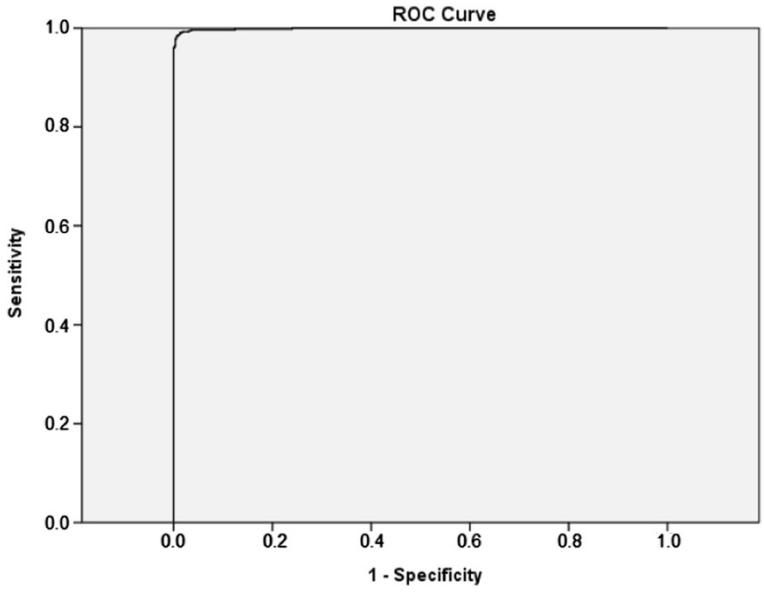

Figure 6 The receiver operating curve (ROC) of the classification between classes in Example 1.

cluster selection method provides a successful separation between the normality and abnormality structures.

In Figure 8, panels A and B show the normal and abnormal synthetic signals in time and TF domains, respectively. The signal on the left-hand side is generated using Equation 1, and the one on the right-hand side is formed by adding three linear FM chirp signals. The TF features extraction and discriminant cluster labeling were applied as explained in the previous example. Figure 8e displays the extracted feature vectors along with the discriminant clusters identified by our proposed method. The features outside this cluster were selected as the commonality structure between the two signals. All the TF features corresponding to the above-selected cluster were chosen, and were used to reconstruct back the TFD. The TFDs corresponding to the common and discriminant structures were plotted in panel D. Observing the ease at which the proposed approach separated the synthetic and chirp-like signal features in this example, it is evident that the this method has the potential to be a powerful and a useful tool in signal pattern recognition applications.

6.2 Real dataset

Pathological voice classification, T-wave alternans (TWA) evaluation from the surface electrocardiogram (ECG), environmental audio classification, telemonitoring of PD are selected as the applications of the developed discriminant cluster selection method. The former is performed using the hard labeling clustering method, and the latter three are evaluated employing the fuzzy labeling approach.

6.2.1 Hard labeling: pathological speech detection

Dysphonia or pathological voice refers to speech problems resulting from damage to or malformation of the speech organs. Currently, patients are required to routinely visit a specialist to follow up their progress. Moreover, the traditional ways to diagnose voice pathology are subjective, and depending on the experience of the specialist, different evaluations can be resulted. Developing an automated technique saves time for both the patients and the specialist, and can improve the accuracy of the assessments. In a previous study from our group [7], we introduced the joint TF feature extraction and classification for pathological speech verification. In this study, we provide this application with a focus on non-stationary TF feature analysis + hard cluster labeling, and compare its performance with traditional clustering methods.

The proposed methodology was applied to the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database, distributed by Kay Elemetrics Corporation [21]. The database consists of 51 normal and 161 pathological speakers whose disorders spanned a variety of organic, neurological, traumatic, and psychogenic factors. The speech signal is sampled at 25 kHz and quantized at a resolution of 16 bits/sample. In this study, 25 abnormal and 25 normal signals were used to train the classifier. Each signal is divided into 80-ms segments and the TFD is constructed [22,23]. Next, NMF with base number of $r = 15$ is employed to each TF representation, and 15 base and coefficient vectors are estimated as explained in Equation (15).

As explained in our previous study [7], abnormal speech behaves differently for voiced (i.e., vowel) and unvoiced

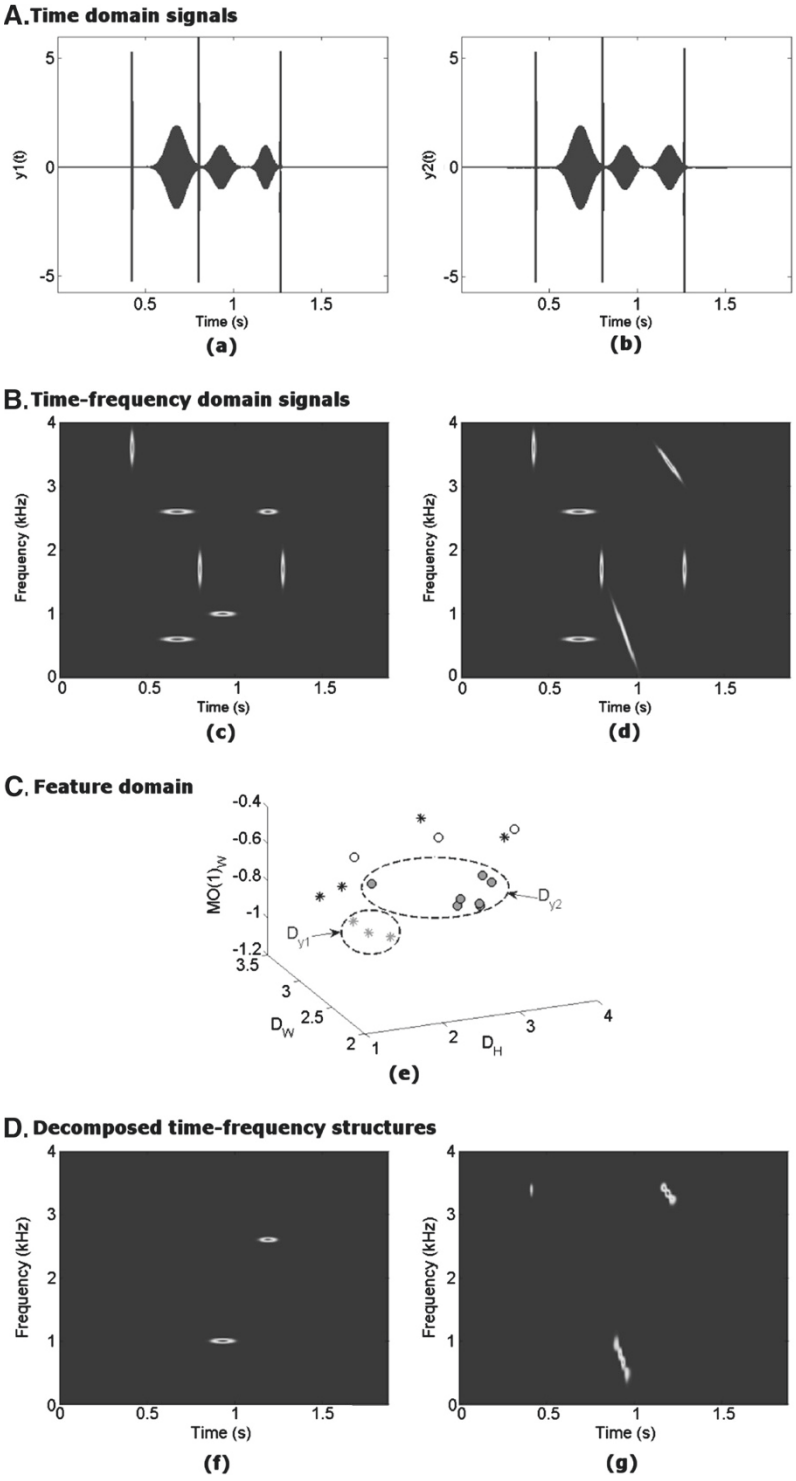

Figure 7 Equation 1 is used to create a synthetic signal y_1 with sampling frequency of 8 kHz: (a) Signal y_1, (α, σ, μ, a) for each component from 1 to 7 is as following: $(6, 0.001, 0.42, 2\pi 3600)$**,** $(1, 0.05, 0.68, 2\pi 2600)$**,** $(1, 0.05, 0.68, 2\pi 600)$**,** $(6, 0.001, 0.80, 2\pi 1700)$**,** $(1, 0.04, 0.93, 2\pi 1000)$**,** $(1, 0.03, 1.18, 2\pi 2600)$**,** $(6, 0.001, 1.27, 2\pi 1700)$**. (b)** Signal y_2 is generated by replacing the 5th and 6th components of signal y_1 with two chirps with slopes of 1 and 0.5 kHz/s, respectively. **(c)** TFD of y_1. **(d)** TFD of y_2. **(e)** TF feature plane including the feature points. Although NMF of order 10 was applied to TF decomposition of each signal, 3 of the y_1 decompositions ($[\vec{w}_i(i)\ \vec{h}_1^T(i)]$) did not have any significant content ($energy_i/energy_{y_1} < 1\%$) so they were excluded from the analysis. D_{y1} and D_{y2} were the clusters identified as the discriminant features in each signal. **(f, g)** The TFD structures selected as the discriminant patterns in y_1 and y_2, respectively.

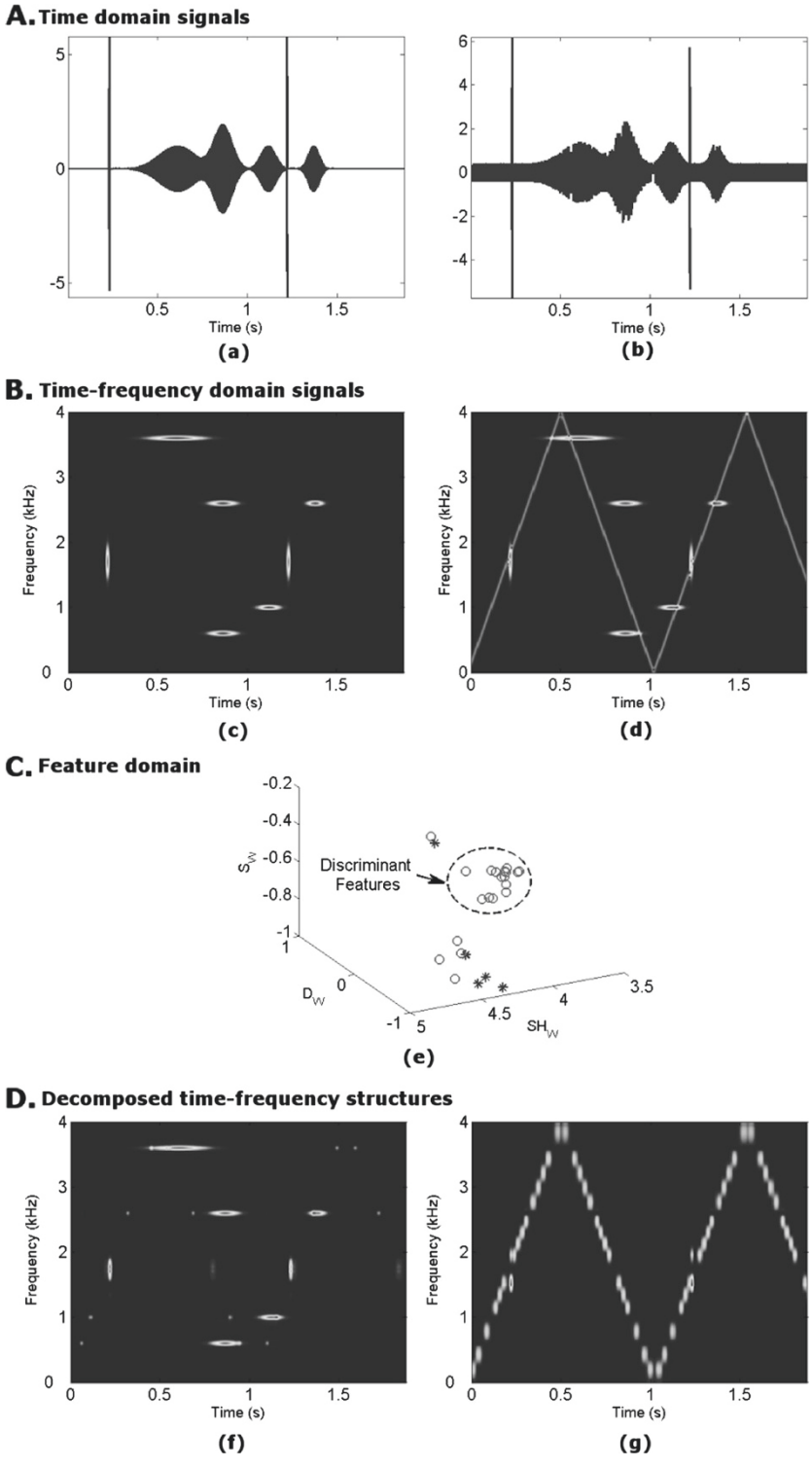

Figure 8 Equation 1 is used to generate a synthetic signal y_1 with sampling frequency of 8 kHz: (a) Signal y_1, (α, σ, μ, a) for each component from 1 to 7 is as following: $(2, 0.1, 0.61, 2\pi 3600)$**,** $(2, 0.05, 0.68, 2\pi 600)$**,** $(2, 0.05, 0.86, 2\pi 600)$**,** $(2, 0.03, 1.37, 2\pi 600)$**,** $(3, 0.001, 0.23, 2\pi 1700)$**,** $(3, 0.001, 1.22, 2\pi 1700)$**,** $(2, 0.04, 1.12, 2\pi 1000)$**. (b)** Signal y_2 is created by adding a chirp to signal y_1. The chirps starts at frequency of $2 * \pi 20$ and reaches to 1 kHz at 0.125 s. **(c)** TFD of y_1. **(d)** TFD of y_2. **(e)** TF feature plane including the feature points. The middle cluster was selected as the discriminant features, and the remaining feature points were identified as the commonality features between the two signals. **(f, g)** the TFD structures selected as the common and discriminant patterns in y_1 and y_2, respectively.

(i.e., constant) components. Therefore, prior to feature extraction, the base vectors are divided into two groups: (a) low frequency (LF): the bases with dominant energy in the frequencies lower than 4 kHz, and (b) high frequency (HF): the bases with major energy concentration in the higher frequencies. Four features ($S_{\vec{h}}, D_{\vec{h}}, S_{\vec{w}}, SH_{\vec{w}}$) are extracted from each LF base vectors, and five features $\left\{S_{\vec{h}}, D_{\vec{h}}, MO_{\vec{w}}^{(1)}, MO_{\vec{w}}^{(2)}, MO_{\vec{w}}^{(3)}\right\}$ are obtained from each HF base vector.

The clustering and labeling are performed as explained in Sections 3.1 and 4.1, respectively. In the training stage, 100 and 20 clusters are experimentally found to be proper choice for the number of clusters (K) in case of LF and HF features, respectively. From the entire clusters, 25% were assigned as common clusters and the remaining clusters labeled class normal or abnormal as explained in hard labeling scheme.

In the test stage, for speech sample, the nearest cluster to each of the TF features are identified using ED criterion shown in Equation (9). Finally, signals with majority of feature vectors in the abnormal clusters are labeled as the pathological speech and the other signals are classified as normal. Figure 9 shows the ROC plot of the proposed TF feature extraction and discriminant cluster selection using hard labeling. The maximum classification accuracy rate of 98.6% is achieved with 50 signals of 51 normal and 159 out of 161 pathological signals are correctly classified. In this figure, the ROC using linear discriminant analysis (LDA) and GMM classifiers are displayed. It can be seen that the proposed discriminant clustering method provides a higher classification accuracy. In [24], well-known Mel-frequency cepstral coefficients (MFCCs) features along with signal pitch is used for pathological speech classification of the same database that we employed in this section. Dibazar et al. [24] achieved an accuracy rate of 98.3% using HMM as the classifier.

6.2.2 Fuzzy labeling: TWA evaluation from the surface ECG

Each year 400,000 North Americans die from sudden cardiac death (SCD). Identifying those patients at risk of SCD remains a formidable challenge. TWA evaluation is emerging as an important tool to risk stratify patients with heart diseases. TWA is a heart rate-dependent phenomenon that manifests on the surface ECG as a change in the shape or amplitude of the T-wave every second heart beat. The presence of large magnitude TWA often presages lethal ventricular arrhythmias. Because the TWA signal is typically in the microvolt range, accurate detection algorithms are required to control for confounding noise and changing physiological conditions (i.e. data non-stationarity). In our previous study [25], we proposed a novel technique, called NMF-adaptive SM [25]. In this method, after pre-processing the ECG recordings to correct baseline wander and removing nonuniform QRS beats [26], the T-wave of each beat is aligned as shown in Figure 10.

Next, the adaptive TFD [22,23] of the aligned T-waves is constructed over each vertical sample ($\vec{A}_1, \vec{A}_2, \ldots, \vec{A}_N$). Adaptive TFD approach is a high-resolution TF representation capable of adaptively tracking non-stationary structures. Adaptive TFD uses the matching pursuit [22] method to decompose the signal over a dictionary of TF atoms. At each iteration, the signal is projected over a dictionary of TF functions and the one which models the greatest fraction of the signal energy is selected.

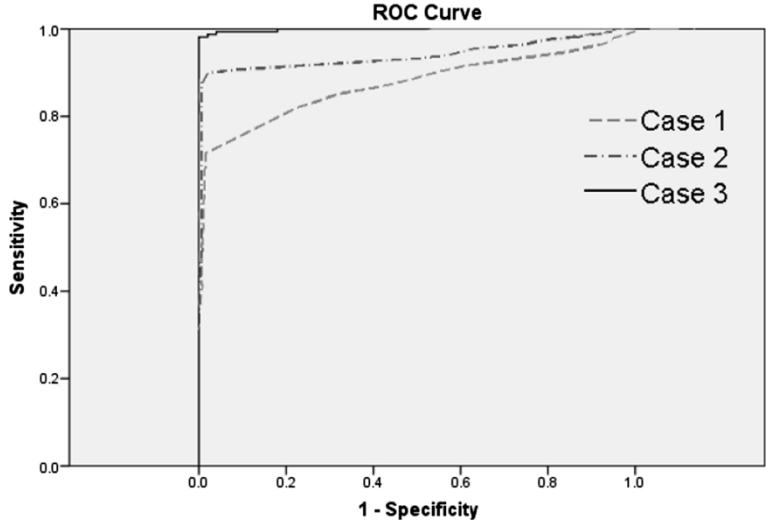

Figure 9 ROC plots for pathological speech classification. Case 1: TF features and LDA classifier. Case 2: TF features and GMM with 15 mixtures. Case 3: TF featured and the K-means clustering along with the proposed hard labeling method.

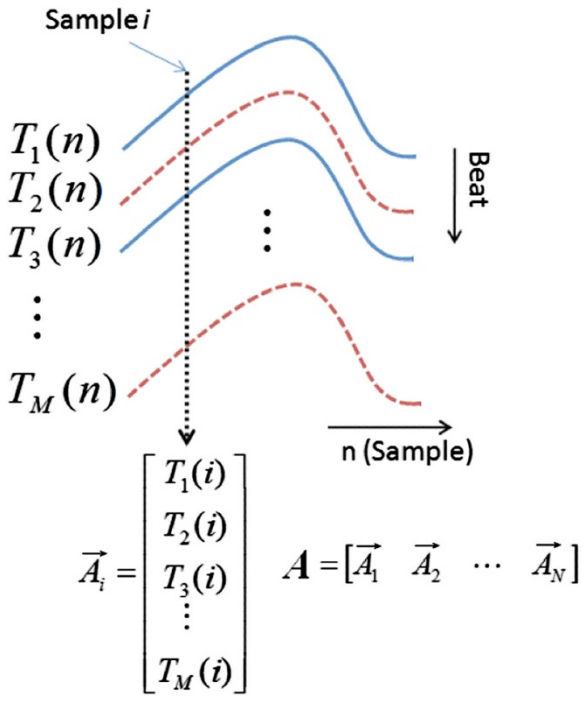

Figure 10 Consecutive T-waves are aligned, and the T-wave amplitude at each sample is transformed into the beat domain.

This TF function is then subtracted from the signal, and the residual signal is subsequently decomposed in further iterations till all or most of the signal energy is modeled. The matching pursuit decomposition with Gabor TF atoms has been chosen in this study because of its superior TF resolution [22], cross-term free nature, adaptivity, and suitability for pattern recognition applications. The adaptive TFD for each vertical sample is computed. If $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_N$ are the TFD of each vertical sample, in the next stage, the TFD representative of the entire T-wave (denoted with $\mathbf{V}_{\text{avg}}$) is calculated as the average of $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_N$. Once the average TFD is constructed, features are extracted as explained below:

NMF with base number of $r = 3$ is employed to each averaged TFD, and three base and coefficient vectors are estimated as explained in Equation (15). NMF is expected to separate the TF structure the noise components that may mask the TWA signal. From each decomposed base component, 11 features are extracted. The first feature is determined as the estimate of the TWA magnitude [27] from each base:

$$a = \text{Real}\left\{\sqrt{T - \mu_{\text{noise}}}\right\} \tag{32}$$

where T is the energy of the decomposed base ($\mathbf{W}$) at frequency of 0.5 cpb (cycle per beat), and μ_{noise} estimates the noise energy. Considering a white Gaussian noise, noise has a constant spectral density at the entire spectral bandwidth. Since the T-wave alternation and respiratory activities do not have any spectral content over the spectral bandwidth of 0.36 to 0.49 cpb, this bandwidth is used to estimate the noise energy. The last ten elements in each base component represent the spectral content of the T-wave signal. Basically, any information about T-wave including noise and TWA value exist in the spectral content of the last ten elements in the base vector. Therefore, the other ten features are chosen to be the last ten elements in each base component. As a result of this feature extraction, 3 feature vectors are extracted for each ECG segment where each vector includes 11 feature values.

As explained in our previous study [25], real-world ECGs with inherent noise were obtained from 26 normal subjects who underwent 2 channel ambulatory ECG recordings (GE Healthcare, Inc.) for 24–48 continuous hours at our institution. The ambulatory ECGs were recorded at a sampling rate of 125 Hz and then exported for custom analysis. The mean heart rate of these recordings was 78–17 bpm (beats per minute) and the mean noise level was 40–67 μV. Each ECG channel was included as a separate record.

Two groups of ambulatory ECGs were generated, one without simulated TWA (TWA magnitude = 0 μV) and the other with simulated TWA (TWA magnitude = 5 μV): ECG signals are recorded from normal subjects and therefore they are assumed to have 0 TWA. A simulated TWA signal with amplitude of 5 μV is added to the ECG signals by uniformly increasing T-wave amplitude of even beats and decreasing T-wave amplitude of odd beats across the T-wave. The use of a known TWA signal permits TWA quantification to be compared between the different methods. A TWA detection threshold of 5 μV was prespecified as this cutpoint approximates the TWA magnitude measured by Klingenheben et al. [28] in patients with heart disease using a similar definition of TWA as our study. The extracted features from NMF-adaptive SM were fed into two classifiers (the SOTM clustering and fuzzy labeling, and an LDA) to train and classify the ECG segments as TWA present or absent.

Half of the dataset is used for the training stage and the other half is employed to test the accuracy of the TWA detection. SOTM is applied on the training dataset and the number of valid clusters is calculated for the classification. Clusters with small number of samples are eliminated. We experimentally decided that the clusters with a membership of less than 1% of the entire feature vectors are not valid. The clusters are formed as the data are presented to the network and the number and size of the clusters is determined by the parameters such as the hierarchical control function ($H(t)$) and the learning rate ($\lambda(t)$). The initial values of these functions are appointed

according to the dataset. In the next stage, the membership coefficients are calculated for each cluster based on the distribution of the train signals. In the test stage, each of the test signals are assigned to one of these cluster centers based on the minimum ED measure. Finally, the class label of each signal is determined by the weighted sum of the feature vectors falling within each cluster multiplied by the membership coefficients. Another point to be discussed here is that since the data are represented to the SOTM in a random manner, the number and the shape and size of the clusters might vary each time the clustering algorithm is run on the data. However, since there is not a one-to-one correspondence between the clusters and the two groups, this fact has no considerable impact on the total performance of the classifier. In addition, the results of the several are averaged to further eliminate this effect.

Table 1 summarizes the results. The proposed TF features and fuzzy labeling classifier (NMF-adaptive SM and fuzzy labeling) were more accurate in detecting the TWA signal than classic LDA classifier. In Table 1, the TWA detection accuracy for our method was further compared with two well-known T-wave analysis methods (spectral methods (SM) [27] and modified moving average (MMA) [26], and two previously described wavelet-based methods (Wavelet 1 [29] and Wavelet 2 [30]). The significant improvement in the sensitivity and specificity of the proposed feature and classifier supports the effectiveness of this approach.

6.2.3 Fuzzy labeling: audio classification

Audio signals are the important sources of information for understanding the content of multimedia. Therefore, developing audio classification techniques that better characterize audio signals plays an essential role in many multimedia applications, such as multimedia indexing and retrieval, and auditory scene analysis. Having approximately 10% of the world population suffering from some sort of hearing loss, one of the important applications of audio classification is in hearing aids (HA) for hearing impaired people. In order to prevent the noise signals from being magnified by the hearing aid, the HA is required to detect the audio classes which the incoming signals belong to, and then change the HAs parameters accordingly. A recent article from our group [31] presented the benefits of joint TF feature extraction employed in environmental audio classification. Next section provides the performance of fuzzy cluster labeling employed along with non-stationary joint TF features when performed for audio scene analysis, and compares its performance with supervised classification.

In this study, we use an environmental audio dataset that was compiled in our signal analysis research group at Ryerson University [31]. The dataset is designed to have 10 different classes such that it consists of 192 audio signals of 5-s duration each with a sampling rate of 22.05 kHz and a resolution of 16 bits/sample. This database is designed to have 10 different classes including 20 aircraft, 17 helicopters, 20 drums, 15 flutes, 20 pianos, 20 animals, 20 birds, and 20 insects, and the speech of 20 males and 20 females. Most of the music samples were collected from the Internet and suitably processed to have uniform sampling frequency and duration.

Three-second audio signals are transformed into TF domain. Next, NMF with decomposition order of 15 ($r = 15$) decomposes each TFD into 15 base and coefficient vectors. In this study, experimentally, $r = 15$ is found to be a suitable choice for the application. Seven features (Section 5) are extracted from each base and coefficient vector. Finally, The SOTM clustering and fuzzy labeling is employed to train and classify the signals.

One of the most important classification tasks for a hearing aid system is to discriminate human speech from environmental noise. Therefore, in the first scenario the dataset consists of signals from human speech and environmental sounds. The human category includes 20 signals from male speakers and 20 signals from female speakers and environmental sounds include 10 bird, 10 aircraft, 10 piano, and 10 animal signals. Table 2 shows the results for this classification task where an accuracy of 96% has been achieved. As it can be seen from the confusion matrices, the system demonstrates high accuracy

Table 1 Comparison of TWA detection rate for NMF-Adaptive SM using the proposed fuzzy labeling classifier and LDA

Method	Sensitivity (%)	Specificity (%)	Classifier	TWA magnitude (μV)	ECG database
NMF-Adaptive SM	92	95	Fuzzy labeling	5	52 Real ECGs
NMF-adaptive SM	87	91	LDA	5	52 Real ECGs
SM [25]	73	63	LDA	5	52 Real ECGs
MMA [25]	92	58	LDA	5	52 Real ECGs
Wavelet 1 [29]	77	NS	Wilcoxon rank-sum test	10	10 Synthetic ECGs
Wavelet 2 [30]	96	NS	LDA	10	2050 Synthetic ECGs

The results of two well-known T-wave analysis methods (SM and MMA) and two wavelet-based TWA detection methods are also reported in the table. 'NS' indicates 'Not Specified'.

Table 2 Confusion matrix for classifying human versus non-human audio signals

Classes		Total	TF + soft labeling		TF + GMM		MFCC + GMM	
			Human	Non-human	Human	Non-human	Human	Non-human
Human	(n)	40	40	0	37	3	40	0
	(%)	100	100	0	92.5	7.5	100	0
Non-human	(n)	40	3	37	8	32	17	23
	(%)	100	7.5	92.5	20	80	42.5	57.5

in discrimination of human voice from other audio signals. The achieved true positive rate shows that all human voice signals have been classified correctly. In addition, the overall accuracy rate for classification scenarios that include discrimination of human voice is very high.

The human versus non-human sound discrimination is also performed using GMM as a successful traditional clustering method for audio signals. This classification resulted in a lower performance with 86% overall accuracy rate. Fifteen mixtures are experimentally found sufficient and used for the GMM classification. We also compared the accuracy of the TF feature extraction and clustering method with the well-known MFCCs features. MFCCs are short-term spectral features and are widely used in the area of audio and speech processing.

In this application, a signal is divided into 32-ms segments and then we compute the first 13 MFCCs for all the segments of the entire length of the audio signals and use them as feature vectors. Using GMM and 15 mixtures, MFCC features resulted in 79% overall classification accuracy rate. It can be seen that MFCC features and GMM system are able to successfully classify human signals; however, the method is not very effective for classifying the non-human signals (i.e., 57.5% accuracy rate). The reason for such behavior can be explained that MFCC features and GMM clustering system are useful for human speech analysis, but they are challenged when dealing with natural sounds with non-human sources. However, it can be evidenced that the combination of the TF feature vectors and the proposed discriminant cluster labeling are significantly successful.

Furthermore, in order to evaluate the efficiency of the system to discriminate human voice in particular environments, two other classification tasks have been defined. In the first case, an accuracy of 98% has been achieved in discrimination of human voice from the musical instruments. This capability could be useful in recognizing and separation of human voice from the background music in a song or at the concert. The second classification task was defined as discrimination of human voice from natural sounds, where an accuracy of 96% has been achieved. Furthermore, the proposed method was applied to other classification scenarios such as natural versus artificial sounds and musical instruments versus aircraft. Table 3 shows the overall obtained average accuracy rate and the dataset used for each classification scenario. The classification accuracy rate using GMM clustering method and the MFCC features are also presented in Table 3.

6.2.4 Fuzzy labeling: telemonitoring of PD

In this application, we present an assessment of the proposed discriminant clustering method for discriminating healthy people from people with PD by detecting dysphonia. The data for this application were obtained from Little et al. [32]. The dataset consists of 195 sustained vowel phonations from 31 male and female subjects, of

Table 3 Different audio classes in the dataset and the number of signals in each class

Classes	Dataset	Average accuracy		
		TF + soft labeling (%)	TF + GMM (%)	MFCC + GMM (%)
Human/non-human	Non-human: aircraft, piano, animal, bird Human: male and female speeches	96	86	79
Human/music	Music: piano, flute, drum Human: male and female speeches	98	68	71
Natural/artificial	Natural: male, female, bird, animal, insect Artificial: helicopter, airplane, piano, flute, drum	91	63	62
Human/Nature	Nature: animal, insect, bird Human: male and female speeches	98	83	75
Aircraft/music	Music: piano, flute, drum Aircraft: helicopter, airplane	98	76	89

which 23 were diagnosed with parkinson disease. The time since diagnoses ranged from 0 to 28 years, and the ages of the subjects ranged from 46 to 85 years (mean 65.8, standard deviation 9.8). Averages of six phonations were recorded from each subject, ranging from 1 to 36 s in length. See [32] for subject details. Little et al. [32] selected ten highly uncorrelated measures, and an exhaustive search of all possible combinations of these measures finds four that in combination lead to overall correct classification performance of 91.4%, using a kernel support vector machine (SVM).

In this section, we employ the ten features proposed in [32] and employ the proposed discriminant clustering method using soft labeling strategy to perform discrimination between people with PD and healthy subjects. It is worth mentioning that since this database provided only the extracted attributes and not the original signals, we could only use the given features. This way, we could evaluate the proposed discriminant cluster selection method and investigate the efficiency of this method in comparison to the exhaustive search and SVM classification used in [32].

Using the proposed discriminant cluster selection and soft labeling method, two common and four discriminant clusters are obtained. This method achieved an overall classification performance of 97% which was higher than 91.4% that was reported in [32]. GMM is also applied for the classification of the PD features. Five and four mixtures are obtained for PD and normal classes, respectively. A poor classification performance with an overall accuracy rate of 69% is obtained using GMM. ROCs for the classification using discriminant clustering and GMM are shown in Figure 11 with the area under the curve of 0.995 and 0.65, respectively.

7 Conclusion

The objective of this article was to improve the performance of pattern recognition systems when there is an overlapping feature vectors due to non-stationarity of the signals or the commonality that exist among different classes. To make this happen, the article introduced a different strategy to clustering techniques based on a fusion of unsupervised and supervised learning approaches. This method applied an unsupervised clustering to the feature vectors from all the different classes, and then used a supervised labeling method to select two types of clusters: discriminant and common clusters. The supervised cluster labeling approach selected the discriminant clusters from the common ones according to their importance for representing each corresponding class. The obtained discriminant clusters represented the differentiating patterns that exist among signals from different classes. Therefore, in the classification stage, only the feature vectors that were located in the discriminant clusters were considered for the classification of a given signal. These feature vectors were better representatives of the signals' characteristics, and resulted in a significantly higher classification accuracy rate.

In order to identify the discriminant clusters, two cluster labeling methods were proposed: hard and fuzzy labeling. In hard labeling, discriminant clusters were assigned to one of the possible classes, but in fuzzy labeling, they were associated to each class with a relative membership value ranging from 0 to 1 (with 0 being the least

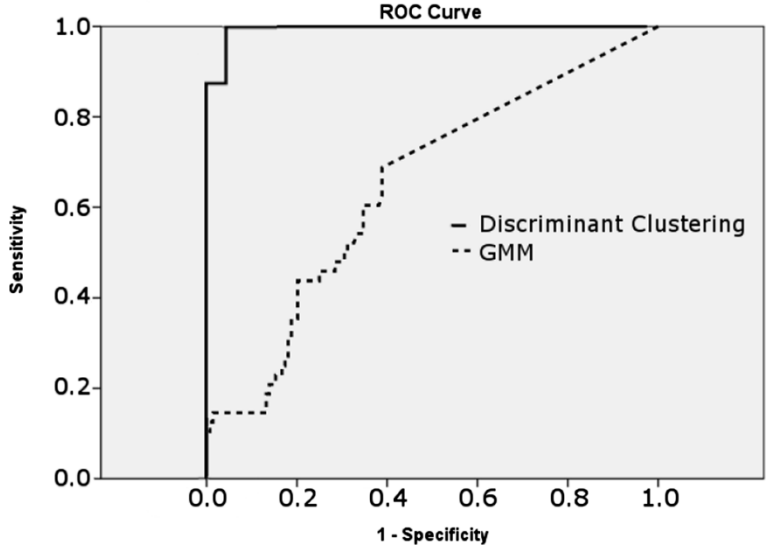

Figure 11 ROC plot for classification of PD by detecting dysphonia. The area under the curve of 0.995 and 0.65 is achieved using the proposed cluster selection method and GMM, respectively.

contribution, and 1 being the most). Both proposed methods enhanced the commonly used supervised learning and clustering approaches. K-means and SOTM clustering methods were explained for the applications studied in this article. An advantage of SOTM compared to the K-means method was the number of clusters, which should be known beforehand in K-means, but was adaptively determined in the SOTM algorithm.

In conclusion, experiments performed with synthetic signals as well as pathological speech, surface ECG, telemonitoring of PD, and environmental audio signals demonstrated the potential of the proposed discriminant feature clustering framework for becoming a powerful pattern recognition tool.

Competing interests
The authors declare that they have no competing interests.

Received: 7 March 2012 Accepted: 4 October 2012
Published: 27 November 2012

References

1. G Freeman, R Dony, S Areibi, in *Proceedings of the IEEE Symposium on Computational Intelligence in Image and Signal Processing, vol. 2846*. Audio environment classication for hearing aids using artificial neural networks with windowed input, (Honolulu, HI, April 2007), pp. 183–188
2. Z Pawlak, Rough sets. Int. J. Comput. Inf. Sci. **11**, 341–356 (1982)
3. Z Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. (Kluwer Academic Publishers, Norwell, 1992)
4. R Jensen, Q Shen, New approaches to fuzzy-rough feature selection. IEEE Trans. Fuzzy Syst. **17**, 824–838 (2009)
5. N Saito, R Coifman, Local discriminant bases and their applications. J. Math. Imag. Vis. **5**(4), 337–358 (1995)
6. S K Umapathy, Krishnan, Time-width versus frequency band mapping of energy distributions. IEEE Tran. Signal Process. **55**, 978–989 (2007)
7. B Ghoraani, S Krishnan, A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification. EURASIP J. Adv. Signal Process. **2009**(ID 928974), 11 (2009). doi:10.1155/2009/928 974
8. A Jain, R Duin, J Mao, Statistical pattern recognition: a review. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 4–37 (2000)
9. R Duda, P Hart, D Stork, *Pattern Classification*. (Wiley, New York, 2001)
10. H Kong, L Guan, in *Proceedings of the IEEE International Conference on Neural Networks,* vol. 2. Detection and removal of impulse noise by a neural network guided adaptive median filter, (Perth, WA, November 1995), pp. 845–849
11. M Kyan, Unsupervised learning through dynamic self-organization: implications for microbiological image analysis. PhD thesis, School of Electrical and Information Engineering University of Sydney, (2007)
12. M Kyan, J Jarrah, P Muneesawang, L Guan, Strategies for unsupervised multimedia processing: self-organizing trees and forests. IEEE Comput. Intell. Mag. **1**, 27–40 (2006)
13. D Lee, H Seung, Learning the parts of objects by non-negative matrix factorization. Nature. **401**(6755), 788–791 (1999)
14. P Paatero, U Tapper, Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. **5**, 111–126 (1994)
15. D Lee, H Seung, in *Advances in Neural Information Processing Systems 13*. Algorithms for non-negative matrix factorization (MIT Press, Cambridge, MA), pp. 556–562
16. C-J Lin, Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)
17. B Tacer, P Loughlin. Time-frequency based classification,in Proceedings of the International Society for Optical Engineering (SPIE), *vol. 2846*, Denver, CO, August 1996), pp. 186–192
18. D Groutage, D Bennink, Feature sets for nonstationary signals derived from moments of the singular value decomposition of cohen-posch (positive time-frequency) distributions. IEEE Trans. Signal Process. **48**(5), 1498–1503 (2000)
19. M Davy, C Doncarli, GF Boudreaux-Bartels, Improved optimization of time-frequency based signal classifiers. IEEE Signal Process. Lett. **8**, 52–57 (2001)
20. M Davy, A Gretton, A Doucet, P Rayner, Optimized support vector machines for nonstationary signal classification. IEEE Signal Process. Lett. **9**(12), 442–445 (2002)
21. M Eye, E Infirmary, *Voice Disorders Database,* Version 1.03. (Kay Elemetrics Corporation, Lincoln Park, 1994)
22. SG Mallat, Z Zhifeng, Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
23. S Krishnan, R Rangayyan, G Bell, C Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology. IEEE Trans. Biomed. Eng. **47**(6), 773–783 (2000)
24. A Dibazar, S Narayanan, T Berger, in *Proceedings of the Second Joint EMBS/BMES Conference,* vol. 1. Feature analysis for automatic detection of pathological speech, (Houston, TX, USA, October 2002), pp. 182–183
25. B Ghoraani, S Krishnan, RJ Selvaraj, VS Chauhan, T wave alternans evaluation using adaptive time-frequency signal analysis and non-negative matrix factorization. Med. Eng. Phys. **33**(6), 700–711 (2011)
26. BD Nearing, RL Verrier, Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy. J. Appl. Physiol. **92**, 541–549 (2002)
27. JM Smith, EA Clancy, CR Valeri, JN Ruskin, RJ Cohen, Electrical alternans and cardiac electrical instability. Circulation. **77**(1), 110–121 (1988)
28. T Klingenheben, P Ptaszynski, S Hohnloser, Quantitative assessment of microvolt t-wave alternans in patients with congestive heart failure. J. Cardiovasc. Electrophysiol. **16**, 620–624 (2005)
29. I Romero, N Grubb, G Clegg, C Robertson, P Addison, J Watson, T-wave alternans found in preventricular tachyarrhythmias in CCU patients using a wavelet transform-based methodology. IEEE Trans. Biomed. Eng. **55**, 2658–2665 (2008)
30. M Boix, B Cantó, D Cuesta, P Micó, Using the wavelet transform for t-wave alternans detection. Math. Comput. Model. **50**, 738–742 (2009)
31. B Ghoraani, S Krishnan, Time-frequency matrix feature extraction and classification of environmental audio signals. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2197–2209 (2011)
32. M Little, P McSharry, S Roberts, D Costello, I Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMed. Eng. **OnLine 6**(23) (2007)

doi:10.1186/1687-6180-2012-250
Cite this article as: Ghoraani and Krishnan: **Discriminant non-stationary signal features' clustering using hard and fuzzy cluster labeling.** *EURASIP Journal on Advances in Signal Processing* 2012 **2012**:250.